Technicolor and INRIA/IRISA at MediaEval 2011: learning temporal modality integration with Bayesian Networks

Cédric Penet, Claire-Hélène Demarty, Guillaume Gravier, Patrick Gros

Technicolor and INRIA/IRISA at MediaEval 2011: learning temporal modality integration with Bayesian Networks*

Cédric Penet, Claire-Hélène Demarty
Technicolor/INRIA Rennes & Technicolor
1 ave de Belle Fontaine
35510 Cesson-Sévigné, France
cedric.penet@technicolor.com
claire-helene.demarty@technicolor.com

Guillaume Gravier, Patrick Gros
CNRS/IRISA & INRIA Rennes
Campus de Beaulieu
35042 Rennes, France
guig@irisa.fr
Patrick.Gros@inria.fr

ABSTRACT
This paper presents the work done in Technicolor and INRIA regarding the Affect Task at MediaEval 2011. This task aims at detecting violent shots in movies. We studied a bayesian network framework, and several ways of introducing temporality and multimodality in the framework.

Keywords
Violence detection, Bayesian Networks

1. INTRODUCTION
The MediaEval 2011 Affect Task aims at detecting violence in movies. A complete description of the task and datasets may be found in [2].

As explained in the paper cited above, violence has not been a widely studied field, despite its importance in parental control. Moreover, most of the work related to violence focuses mainly on either video or audio. However, in movies, our intuition is that every information channel is important in order to correctly assess violence, even for a human assessor, and that these information streams are complementary. This paper presents our results for both video and audio only systems and for the fusion of both sources of information. We also investigate the importance of temporality in the framework.

2. SYSTEM DESCRIPTION
The system we developed for the task contains four different parts:

2.1 Features extraction
Movies have several components that convey different information, namely audio, video, subtitles, ... We chose to extract features from the audio and video modalities.

Audio modality Five audio features were extracted from the audio stream from 40 msec frames with 20 msec overlap. These features were then averaged over video shots, as the task is to be performed at the shot level. The audio features are the energy, the centroid, the asymmetry, the zero crossing rate (ZCR) and the flatness.

*This work was partly achieved as part of the Quaero Program, funded by OSEO, French State agency for innovation.

MediaEval 2010 Workshop, September 1-2, 2011, Pisa, Italy

Video modality Four video features were extracted from the video for each shot: the shot duration, the average number of blood pixels, the average activity and the number of flashes.

The features histograms were then rank-normalised over 22 values for each movie.

2.2 Classification
In our process, we chose to use Bayesian networks (BN) [3] as a probability distribution modelisation process. BN exhibit interesting features: they model statistical distribution using a graph whose structure may be learned. On the downside, they consider each feature independently.

For the MediaEval 2011 Affect Task, several types of structures were tested:

Naive Bayesian network (NB) This is the simplest network. Each feature is only linked to the classification node. It therefore makes the assumption that the features are all independent with respect to the classification node.

Forest augmented naive Bayesian network (FAN) This structure has been introduced in [4]. The idea here is to relax the assumption of features independence with respect to the classification node. The algorithm learns a forest between the features before connecting them to the classification node.

K2 [1] This structure learning algorithm is a state-of-the-art score based greedy search algorithm. In order to reduce the number of possible graphs to test, it requires a nodes ordering.

We used the Bayes Net Toolbox[1].

2.3 Temporal integration
Considering the temporal structure of movies we decided to try several types of temporal integrations: we used contextual features[2] over $n \in [-5, +5]$, and two types of temporal filtering over 5 samples, that are used to smooth decisions:

decision maximum vote This intervenes once the decision has been taken and consists in taking the maximum decision over a few samples.

[1] `http://code.google.com/p/bnt/`

[2] Considering $X_t = [x_1, \cdots, x_K]$ the feature vector for sample at time t, the contextual features vector becomes $X_t^c = [X_{t-n}, \cdots, X_t, \cdots, X_{t+n}]$.

Description						MC	F
LF	C	A: Me	V: Ma	A: N	V: K2	**0.761**	**0.397**
LF	C	A: Me	V: Me	A: N	V: K2	**0.774**	**0.391**
V	C	Ma		K2		**0.784**	**0.305**
A	C	Me		K2		**0.805**	**0.295**
V	C	Me		K2		**0.840**	**0.297**
A	C	Me		N		**0.843**	**0.354**
EF	C	Ma		K2		**0.892**	**0.284**
A	C	Ma		K2		**0.943**	**0.268**
V	-	Ma		N		**0.950**	**0.255**
A	-	-		K2		**0.967**	**0.251**
EF	C	-		K2		**0.998**	**0.266**
V	-	Ma		FAN		**1.009**	**0.276**

Table 1: Results for runs submitted ordered by increasing value of MediaEval Cost (MC) (F: F-measure, LF: late fusion, EF: early fusion, A: audio, V: video, C: contextual, N: naive BN, Ma: max decision vote, Me: mean probability).

probability averaging This intervenes before taking the decision, by directly averaging the samples probabilities of being violent.

2.4 Modalities fusion

As for multimodal fusion, two cases were considered: late fusion and early fusion. For early fusion, we simply fused the features from both modalities before learning, while for late fusion, we fused the probability of both modalities for the ith shot s_i using:

$$P^{s_i}_{fused}(P^{s_i}_{v_a}, P^{s_i}_{v_v}) = \begin{cases} max(P^{s_i}_{v_a}, P^{s_i}_{v_v}) & \text{if both are violent} \\ min(P^{s_i}_{v_a}, P^{s_i}_{v_v}) & \text{if both are non violent} \\ P^{s_i}_{v_a} * P^{s_i}_{v_v} & \text{otherwise} \end{cases} \quad (1)$$

where $P^{s_i}_{v_a}$ (respectively $P^{s_i}_{v_v}$) is the probability of being violent for the audio (respectively video) modality for the ith shot.

This rule gives high scores when both audio and video find a violent segment, a low score if they both do not and an intermediate score if only one answers yes.

3. RUNS SUBMITTED AND RESULTS

This section describes the runs submitted to the MediaEval 2011 Affect Task. For the audio and video experiments, we chose to submit the two best runs according to the MediaEval cost and to the false alarm vs missed curve using cross validation on the learning set, while for the multimodal runs (namely, early and late fusion), we chose the best ones according to both metrics. The selected runs and their results are presented in table 1.

Most of the obtained scores have values lower than < 1 which is better than the simple case where each sample is classified as violent, i.e. false alarm rate is 100% and missed detection rate is 0%.

The analysis of the produced graphs yields nice and encouraging observations on the quality of the structure learning algorithms. Firstly, the links between features may be easily interpreted. The ZCR and centroid are linked as they represent the same information, the activity is linked to the shot length as the shot detector used tends to oversegment when the activity is high, and finally blood is not connected to violence, which seems logical considering that the presence of blood in the violent scenes highly depends on the movie and the chosen definition for violence. As for EF, while we thought it would improve the results and find correlations between audio and video, it seems that for non contextual data, only the video features are linked to the violence node, and that for contextual data the links are messy. This and the better results obtained using LF tend to indicate that the features used in both modalities are from different levels and cannot be compared as such. Secondly, it seems that the algorithms produce a strict temporal structure, i.e. the features from time $t = n$ are linked together and not to features from different times unless they are in chains. There are four chains in the graphs: flatness, energy, activity and blood. It is easy to see that these features have a temporal structure. On the other hand, the flash feature is connected only to the violence node and forms no chain, which is again logical as the flash feature only detects high luminance variations, and has therefore no well definite temporal structure.

The use of contextual seems to provide good and promising results, which tends to confirm the importance of the temporal structure of movies. The depth used for this evaluation has been chosen arbitrarily, however it should be interesting to also consider other depths. On the downside, it seems that these results depend on the algorithm used for learning the BN structures: FAN and non contextual data seem to work better, while K2 and contextual data seem to give the best results.

This concludes the preliminary analysis that can be inferred from the evaluation.

4. CONCLUSION

This paper presents a simple framework based on temporal integration, multimodality and Bayesian network. First, it is experimentally shown that the structure learning algorithm output logical graph with respect to the provided data: they are able to capture the links between features and provide a coherent temporal structure. It is also shown that early fusion with features that have different nature yields to poor results, while late fusion seems to be more promising. Second, the use of contextual data seems to improve the result.

This work provides a promising baseline for future work on the subject. We have several improvement ideas. We want to add features from the text modality, as we think it also contains important information on the violent nature of the video shots. We also want to investigate more the contextual data and test other structure learning algorithms.

5. REFERENCES

[1] G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[2] C.-H. Demarty, C. Penet, G. Gravier, and M. Soleymani. The MediaEval 2011 Affect Task: Violent Scenes Detection in Hollywood Movies. In *MediaEval 2011 Workshop*, Pisa, Italy, September 1-2 2011.

[3] D. Heckerman. A Tutorial on Learning with Bayesian Networks. Technical report, Microsoft Research, 1995.

[4] P. Lucas. Restricted Bayesian Network Structure Learning. In *Advances in Bayesian Networks, Studies in Fuzziness and Soft Computing*, pages 217–232, 2002.